Índices y Operaciones para Bases de Datos Métricas

M. D. Alba, J. Arroyuelo, M. E. Di Genaro, A. Grosso, M. Jofré, V. Ludueña, N. Reyes
Dpto. de Informática, Fac. de Cs. Físico-Matemáticas y Naturales, Universidad Nacional de San Luis
{*bjarroyu, digeme, agrosso, vlud, nreyes*}@unsl.edu.ar,{*mdaniela.alba, monicajofre*}@gmail.com
Edgar Chávez
Centro de Investigación Científica y de Educación Superior de Ensenada, México
elchavez@cicese.mx
Karina Figueroa
Fac. de Cs. Físico-Matemáticas, Universidad Michoacana de San Nicolás de Hidalgo, México
karina@fismat.umich.mx
Rodrigo Paredes
Dpto. de Cs. de la Computación, Fac. de Ingeniería, Universidad de Talca, Chile
raparede@utalca.cl

Resumen

La disponibilidad de dispositivos electrónicos en diversos ámbitos y al alcance de todos, junto con el uso masivo de las redes, ha provocado que una gran cantidad y variedad de datos sean generados cada segundo. Este contexto motivó que las bases de datos debieran adaptarse a nuevos tipos de datos (no estructurados) y evolucionaran para lograr administrar de manera eficiente ese gran volumen de datos, al igual que el tipo de requerimientos al que son sometidos los mismos.

Un modelo de base de datos que se adapta al entorno descrito son las *Bases de Datos Métricas*. Esta investigación pretende contribuir a la madurez de este modelo de bases de datos, considerando distintas perspectivas como la administración del espacio disponible (crucial debido a la gran cantidad de datos); formas más sofisticadas de búsqueda sobre las mismas; optimización de estos depósitos, o desarrollo de nuevos, considerando incluso la arquitectura del procesador, entre otros.

Palabras Claves: bases de datos métricas, índices, búsquedas por similitud.

Contexto

Esta investigación se realiza en el marco del Proyecto Consolidado *Tecnologías Avanzadas de Bases de Datos* de la Universidad Nacional de San Luis (Código 03-2218) y en Programa de Incentivos (Código 22-F814), dentro de la línea *Bases de Datos no Convencionales*. Colaboran investigadores de otros grupos de la región: Universidad de Talca (Chile), Universidad Michoacana de San Nicolás de Hidalgo y Centro de Investigación Científica y de Educación Superior de Ensenada (México). Mediante RCS N° 186/2020 se prorrogaron los proyectos de la UNSL hasta diciembre de 2022.

Esta línea tiene como principal objetivo lograr la consolidación del modelo de Bases de Datos Métricas para soportar distintos tipos de bases de datos no convencionales Para este propósito, se está investigando sobre obtener índices que sean escalables, capaces de manejar grandes volúmenes de datos, sin degradar significativamente su desempeño; con operaciones de E/S eficientes; y que además de ser dinámicos, permitan tanto inserciones como eliminaciones, resulten más eficaces considerando el nivel de la jerarquía de memorias en el que trabajen. Igualmente, se analizan nuevas arquitecturas del procesador en un intento de obtener mejoras, a un muy bajo nivel, en los administradores de estas bases de datos. Se espera, de esta manera, contribuir en diferentes campos de aplicación: sistemas de información geográfica, robótica, visión artificial, diseño asistido por computadora, computación móvil; entre otros.

Introducción

El acceso generalizado a dispositivos capaces de generar datos digitales en ámbitos tan diferentes como el educativo, productivo, laboral, de la salud, recreativo, científico, etc. ha ocasionado que el volumen de datos generados y almacenados y la variedad

de sus tipos de datos, crezca de manera exponencial. Como respuesta, las bases de datos han debido adaptarse rápidamente tanto a la cantidad de datos, que deben ser administrados eficientemente, como a su variedad y disimilitud. En este contexto los repositorios especializados en datos *no estructurados* se vuelven indispensables.

El tipo de requerimientos al que son sometidos estos datos también es variado, dado que los mismos pueden provenir de entornos muy diferentes; por ejemplo, se puede ingresar una melodía esperando encontrar canciones semejantes a dicho trozo, o buscar las huellas digitales más similares a una dada; en estos casos las búsquedas tradicionales (exactas) carecen de sentido, en cambio las *búsquedas por similitud* resultan más apropiadas. En este tipo de búsqueda se suele dar un objeto como modelo de lo que se quiere recuperar y se busca en la base de datos los objetos que sean suficientemente similares a él. Estas búsquedas se las conoce como consultas por contenido o consultas mediante un ejemplo (query by example).

Un modelo que resulta adecuado, por englobar muchas de las características compartidas por estas necesidades tan diversas, es el de *espacios métricos*. Este modelo es determinado por un universo de objetos y una función de distancia definida entre ellos, que mide cuán diferentes son. Cualquier tipo de objetos no estructurados, que admita la definición de una medida que para un par de elementos indique cuán diferentes son los mismos, admite ser modelizado de esta manera. La única restricción es que esa medida cumpla con las propiedades que la hagan una métrica. En general, esas medidas sobre conjuntos particulares de datos son provistas por expertos (por ejemplo: distancias que sirven para comparar huellas dactilares).

Además, para responder eficientemente a requerimientos tan dispares sobre estas bases de datos, sin realizar una examinación secuencial del conjunto de datos, son necesarios los llamados *Índices Métricos* o *Métodos de Acceso Métricos* (MAMs). La actualización y optimización de estos índices se vuelve esencial cuando se considera la variedad de ámbitos en los que se aplica este modelo (reconocimiento de voz, reconocimiento facial, bases de datos médicas, minería de datos, biología computacional, etc.), para lograr su adaptación a cada caso particular y afrontar retos como el soporte de conjuntos masivos de datos, el permitir actualizaciones (inserciones/eliminaciones), la resolución de búsquedas más complejas y mejorar el desempeño de los administradores de bases de datos (DBMS) también a bajo nivel.

Líneas de Investigación y Desarrollo

Bases de Datos no Convencionales

En este ámbito de investigación, las bases de datos que administran vídeos, imágenes, texto libre, secuencias de ADN o de proteínas, audio, etc., las llamadas *bases de datos no convencionales*, serán modelizadas utilizando el modelo de espacios métricos, por lo que también son referenciadas como *bases de datos métricas*.

Para responder eficientemente a los distintos requerimientos sobre las mismas, evitando comparaciones exhaustivas sobre la base de datos durante una consulta por similitud, se requiere el uso de índices especializados. Por ello, un objetivo prioritario es optimizar los mismos, analizando aquellos que han mostrado buen desempeño en las búsquedas para reducir su complejidad considerando, cuando sea necesario, el nivel de la memoria en lo que se lo alojará y/o priorizando además,cuando sea posible, su dinamismo y escalabilidad. Debido a lo costoso que suele resultar calcular la distancia entre dos objetos, el número de cálculos realizados al crear el índice o al realizar búsquedas, es usado como medida general de complejidad y será el parámetro que se debe optimizar.

Formalmente, un espacio métrico está definido por un universo de objetos $\mathbb{U}$ y una función de distancia entre ellos $d : \mathbb{U} \times \mathbb{U} \longmapsto \mathbb{R}^+$, que permite medir la disimilitud entre los objetos del universo. En particular, d cumple con las propiedades de una métrica (reflexividad, positividad estricta, simetría y desigualdad triangular), lo que resulta muy útil al momento de resolver consultas por similitud. En particular, la propiedades que permiten ahorrar cálculos de distancia son la simetría y la desigualdad triangular. Una base de datos X es un conjunto $X \subseteq \mathbb{U}$ y una consulta por similitud sobre la misma, $q \in \mathbb{U}$, en general suele ser de dos tipos: *por rango* (q, r) o de *k-vecinos más cercanos* (k-NN(q)).

Búsqueda de los k Vecinos

El localizar las estaciones de servicio más cercanas a nuestra ubicación, es un requerimiento que puede resolverse en espacios métricos mediante una consulta k-NN. Este tipo de consulta resulta muy útil

en una variedad de aplicaciones como la predicción de funciones, el aprendizaje automático, la cuantificación y compresión de imágenes, etc. Una variación menos utilizada de la misma es la que permite obtener los k-vecinos más cercanos de *todos* los elementos de una base de datos, la *All-k*-NN, que resulta tan útil como la anterior. El planteo general de esta consulta relaciona cada elemento $u \in X$, con los k objetos en X- $\{u\}$ que tengan la menor distancia a él.

La solución ingenua tiene una complejidad de $O(n^2)$ cálculos de distancia, si $|X| = n$, y consiste en comparar cada objeto en la base de datos con todos los demás, para obtener la respuesta a esta consulta. Evidentemente, es necesario proceder de manera mucho más eficiente, por ejemplo a través de un preprocesamiento la base de datos; así se construye un índice, y luego se buscan en él los k-NN de cada elemento del conjunto.

Sin embargo, en algunas situaciones esta solución puede resultar tan ineficiente como la ingenua. Esto ocurre cuando se trabaja con espacios métricos de alta dimensión, o si la función de distancia utilizada es demasiado costosa de calcular, o cuando se administra una base de datos masiva. Estas circunstancias pueden requerir revisar la base de datos completa, sin importar la estrategia implementada. Por otro lado, al considerar los requerimientos de algunas aplicaciones particulares, hay algunos que priorizan la velocidad de respuesta sobre la precisión de la misma [12, 6, 13, 7]. En este contexto, toman preponderancia las llamadas *búsquedas por similitud aproximadas*, que admiten mejorar la velocidad de la respuesta a las consultas a costa de aceptar algunos "errores" en la respuesta.

El *Grafo de los k-vecinos más cercanos* (kNNG) [11] se encuentra entre las soluciones propuestas para espacios métricos generales. Este grafo asocia cada elemento de la base de datos a sus k vecinos más cercanos y resulta ser un índice eficiente, con respecto a algunas de las técnicas clásicas. Sabiendo que resolver el problema de los *All-k*-NN permite construir el kNNG, se han propuesto nuevas técnicas, que responden al problema de *All-k*-NN, y permiten computar una aproximación del kNNG. Éstas, conectan cada objeto u de la base de datos con k vecinos *cercanos*, relajando la condición que exige que no haya, en toda la base de datos, algún objeto más cercano a u que los k vecinos devueltos. Aunque estas técnicas pueden perder algún objeto muy cercano a u y en su lugar devolver otro un poco más lejano obtienen, a cambio, una respuesta más rápida. Al grafo construido de esta manera se lo denomina *Grafo de vecinos cercanos* (knNG) [4]. Una característica común en estas propuestas es que ninguna resuelve el problema a través de *buscar en un índice*.

El profundo conocimiento que se tiene del *Árbol de Aproximación Espacial DistalDiSAT* [5] permite plantear una primera aproximación que propone un enfoque ingenioso para un caso particular del problema, se resuelve el *All*-1-NN. Esta técnica utiliza información provista por la *construcción* del *DiSAT* para obtener el 1nNG, conectando a cada objeto con un elemento *cercano* de la base de datos, que puede ser, o no, su vecino más cercano [4]. Esta propuesta permite recuperar el 1nNG con bajo costo, logrando muy buena precisión, un error bajo y por ende un buen compromiso calidad/tiempo, y *sin realizar ninguna búsqueda en el índice*.

Las demás propuestas desarrolladas no se apoyan en ningún índice y se enfocan en el problema general, resolviendo la consulta *All-k*-nN y computando el knNG. La base de estos desarrollos es aprovechar de manera ingeniosa las propiedades que cumple la *función de distancia*. En ellas se sugieren distintas maneras de seleccionar muestras de la base de datos, a partir de las cuales se obtiene un conjunto de distancias que serán el punto de partida de este proceso. Se analizan diferentes maneras de utilizar la información conseguida para calcular, en algunos casos, los vecinos exactos [3] y en otros los aproximados para todos los objetos de la base de datos, utilizando las propiedades mencionadas previamente de simetría o desigualdad triangular.

Índices Métricos

Como se ha explicado, los índices métricos o MAMs permiten responder a las búsquedas sobre conjuntos de datos no estructurados de manera mucho más eficiente que si se examina toda la base de datos de manera secuencial [6]. Para hacerlo los índices métricos aprovechan tanto las distancias almacenadas en él, cómo el hecho de que la función de distancia satisface la simetría y la desigualdad triangular, ahorrando de esta forma cálculos de distancia y tiempo.

Teniendo en cuenta el variado espectro de aplicación de los índices métricos, al momento de estudiar su optimización se deben considerar diferentes puntos de vistas; como tener en cuenta su dinamismo, analizar si se adaptan adecuadamente al almacenamiento en memoria secundaria, considerando como

medida de complejidad en ese caso no sólo el número de cálculos de distancia, sino también el número de operaciones de E/S necesarias, entre otros.

El desarrollo del *Árbol de Aproximación Espacial Dinámico (DSAT)* [10], logró proveer dinamismos a uno de los índices de mejor desempeño en espacios métricos de mediana a alta dimensión, el *Árbol de Aproximación Espacial (SAT)*, el cual a pesar de su eficiencia era totalmente estático. El *DSAT* permite realizar inserciones y eliminaciones; es decir, se puede crear incrementalmente, manteniendo el buen desempeño en las búsquedas. Sin embargo, el DSAT agrega un parámetro a sintonizar, mientras que el SAT carecía de parámetros.

Siguiendo con el enfoque del dinamismo, se propuso en primera instancia el desarrollo de la *Foresta de Aproximación Espacial Distal (DiSAF)* [2]. Este índice pretendía dinamizar el *Árbol de Aproximación Espacial Distal (DiSAT)*, que además de no necesitar ningún parámetro extra, lograba mejorar las búsquedas respecto de sus antecesores SAT y DSAT, sin embargo también era estático. La *DiSAF* es dinámica y aplica la técnica de dinamización de Bentley y Saxe al *DiSAT*. Además, aprovecha el profundo conocimiento que se tiene sobre la aproximación espacial para lograr mejorar al máximo su desempeño. Este índice se diseñó para memoria principal, sin embargo, los costos de construcción son altos debido a la necesidad de reconstruir subárboles luego de cada inserción. Como consecuencia del análisis de los motivos que hacían costosa a la *DiSAF*, principalmente en su construcción, se diseñó el *Árbol de Aproximación Espacial Distal Dinámico (DDiSAT)*. Esta versión utiliza *inserción perezosa* para amortizar los costos de reconstrucción, mientras se mantiene su desempeño en las búsquedas. En la actualidad se están evaluando sus resultados sobre diferentes bases de datos métricas.

Otro aspecto a considerar es cuando los índices no caben en memoria principal, ya sea porque administran una base de datos masiva, o porque los objetos almacenados en la misma son muy grandes (imágenes satelitales). Entonces surge la necesidad de diseñar índices que utilicen adecuadamente la memoria secundaria. En este contexto, se está trabajando en lograr una versión dinámica del *DDiSAT* para memoria secundaria, que además de amortizar los costos de reconstrucción entre varias inserciones y mantener un buen desempeño en las búsquedas, se adapte a memoria secundaria realizando un buen uso de las páginas de disco, a fin de minimizar el número de operaciones de E/S. Para lograrlo, se debe considerar no sólo que logre un desempeño comparable al de la versión de memoria principal en cantidad de cálculos de distancia, sino que las páginas en disco mantengan una buena ocupación, que la cantidad de operaciones de E/S necesarias sea baja y que se mantenga la localidad en los accesos.

Los requerimientos de aplicaciones que priorizan la rapidez en las respuestas, a costa de perder algunos elementos de la misma, es una faceta tan importante como las consideradas anteriormente. A este tipo de búsquedas, en las que se intercambia precisión (devolviendo sólo algunos objetos relevantes) por velocidad en la respuesta (esos objetos se devuelven más rápido), se denominan *aproximadas*. Para conjuntos de datos masivos, las búsquedas por similitud aproximadas permiten obtener un buen balance entre el costo de las búsquedas y la calidad de la respuesta obtenida.

El diseño de la *Lista Dinámica de Permutaciones Agrupadas (DLCP)* [8], que además de ser dinámica es para memoria secundaria, fue diseñada usando como base uno de los mejores representantes de este tipo de consultas, el *Algoritmo Basado en Permutaciones (PBA)* [1], que logra una respuesta de alta calidad a un bajo costo. La *DLCP*, que combina *LC* con *PBA*, agrupa por distancia entre las *permutaciones* de los objetos, en lugar de hacerlo por distancia entre objetos, y además se le puede indicar cuántos cálculos de distancia y/o operaciones de E/S utilizar, para obtener una respuesta rápida, aunque menos precisa. Actualmente, se está considerando también una versión de *DiSAF* para memoria secundaria, que mantenga el dinamismo en la construcción y que permita aprovechar la información de los *DiSAT* que la integran y en la cual las búsquedas por similitud sean aproximadas, para lograr reducir la cantidad de accesos a disco y la cantidad de cálculos de distancia.

Resultados y Objetivos

Los resultados obtenidos en los estudios sobre el modelo de espacios métricos además de lograr mejorar el desempeño de los índices métricos analizados, conducen a estudiar su aplicación a otros métodos de acceso [3, 4, 9, 5, 10, 2], lo que se está llevando adelante como distintas tesis de posgrado.

Se espera brindar nuevas herramientas, que administren bases de datos métricas, de manera de acercar a estas bases de datos no convencionales a la

madurez de los modelos tradicionales. Por tal motivo, se continuará el estudio de nuevos diseños de estructuras de datos, que se adapten tanto al nivel de la jerarquía de memorias donde se almacenarán, como a las características de los datos a indexar, con el fin de mejorar su eficiencia en tiempo y espacio. Del mismo modo, se continuará indagando sobre técnicas innovadoras que, sin utilizar índices, permitan resolver consultas eficientemente. Además, se espera mejorar el desempeño de las operaciones de bajo nivel en los DBMS, mediante una nueva arquitectura del procesador.

Actividades de Formación

Dentro de esta línea de investigación se forman alumnos y docentes-investigadores participando en:

- **Maestría en Cs. de la Computación** (UNSL): tesis sobre una versión dinámica eficiente del *DiSAT*.
- **Maestría en Informática** (UNSJ): tesis sobre la evaluación del knNG para búsquedas por similitud.
- **Maestría en Informática** (UNSJ): tesis sobre una versión dinámica y para memoria secundaria del *DDiSAT*

Referencias

[1] E. Chávez, K. Figueroa, and G. Navarro. Effective proximity retrieval by ordering permutations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(9):1647–1658, Sept 2008.

[2] E. Chávez, M. Di Genaro, N. Reyes, and P. Roggero. Decomposability of disat for index dynamization. *Computer Science & Technology*, pages 110–116, 2017.

[3] E. Chávez, V. Ludueña, and N. Reyes. Solving all-k-nearest neighbor problem without an index. In *Procs. del XXV Congreso Argentino de Ciencias de la Computación CACIC 2019)*, pages 567–576. UniRío editora, 2019.

[4] E. Chávez, V. Ludueña, N. Reyes, and F. Kasián. All near neighbor graph without searching. *Computer Science & Technology*, 18:61–67, 2018.

[5] E. Chávez, V. Ludueña, N. Reyes, and P. Roggero. Faster proximity searching with the distal {SAT}. *Information Systems*, 59:15 – 47, 2016.

[6] E. Chávez, G. Navarro, R. Baeza-Yates, and J. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.

[7] P. Ciaccia and M. Patella. Approximate and probabilistic methods. *SIGSPATIAL Special*, 2(2):16–19, 2010.

[8] K. Figueroa, C. Martínez, R. Paredes, N. Reyes, and P. Roggero. Dynamic list of clustered permutations on disk. In *Computer Science and Technology Series: XXI Argentine Congress of Computer Science Selected Papers*, pages 201–211. EDULP, 2016.

[9] K. Figueroa, N. Reyes, A. Camarena-Ibarrola, and L. Valero-Elizondo. Improving the list of clustered permutation on metric spaces for similarity searching on secondary memory. In *10th Mexican Conference on Pattern Recognition (MCPR2018)*, volume 10880, pages 82–92, 2018.

[10] G. Navarro and N. Reyes. New dynamic metric indices for secondary memory. *Information Systems*, 59:48 – 78, 2016.

[11] R. Paredes, E. Chávez, K. Figueroa, and G. Navarro. Practical construction of k-nearest neighbor graphs in metric spaces. In *Proc. 5th Workshop on Efficient and Experimental Algorithms (WEA)*, LNCS 4007, pages 85–97, 2006.

[12] H. Samet. *Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006.

[13] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach (Advances in Database Systems)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. XVIII, 220 p., Hardcover ISBN: 0-387-29146-6.